LETTER-DISTRIBUTIONS OF WORDS

A. ROSS ECKLER
Morristown, New Jersey

For many years, computational linguists have studied the statistical behavior of language -- the distribution of the number of letters in words, the distribution of the number of words in sentences, etc., in English-language texts. Similarly, cryptanalysts have long been interested in the distribution of the different letters as they occur in English-language text, in order to aid them in decoding substitution ciphers. Claude Shannon and others, in developing the science of information theory during the 1940s, attempted to find out to what extent the properties of English-language text can be modeled by a random process in which the choice of a letter depends upon which letters immediately precede it. (It is a well-known fact that if the letter Q is encountered in text, the odds are very high that the next letter will be U; English abounds with analogous, although less extreme, dependences between nearby letters.)

These statistical studies concentrate on either the large-scale or the small-scale properties of words -- that is, their length or the precise letters (or letter-orderings) they contain. Little attention has been given to the statistical properties of words at an intermediate level. At least two such levels of word study can be identified: the patterns of letters in words, and the distribution of letters in words. In studying letter-patterns, one is interested in whether or not letters in a word are the same, but not in what these letters are; for example, EXCESS and BAMBOO have identical letter-patterns, even though they have no letters in common. In studying letter-distributions, one is interested in the number of letters of different types in a word, but not in the arrangement of these letters in a word; for example, INTONATION and OPPRESSORS both have three repetitions of one letter, two repetitions of three more letters, and one repetition of a fifth letter. This article is concerned solely with letter-distributions in words; an analogous study of letter-patterns is a more formidable undertaking, and can be done best on a sample basis (for example, a study of the 210 different patterns of eight-letter words with two pairs of like letters).

The study of letter-patterns and letter-distributions in English words has been greatly aided by the 1971-73 publication by Jack Levine of a three-volume computer printout of approximately 442,000 words from Webster's Second and Third Unabridged (including inferred plurals of nouns, past tenses of verbs, and gerunds) grouped together by letter-pattern. (For details of this work, now unfortuna-

tely out of print, see the February 1972, November 1972 and August 1973 issues of Word Ways.) Mr. Levine, a professor of mathematics at the University of North Carolina, earlier issued a booklet entitled A List of Words Containing No Repeated Letters, recently available from the American Cryptogram Association (see the May 1972 Word Ways). These words are not included in his three-volume opus, and in fact are based on a different set of dictionaries. As will be shown later, this creates difficulties when comparing the statistical properties of his word lists with other word lists.

To facilitate discussion, one must introduce a shorthand notation to describe the different possible letter-distributions of words. A given distribution will be identified by an increasing sequence of integers specifying how many times each different letter is repeated; for example, INTONATION and OPPRESSORS (see above) both have the letter-distribution 12223. When the word-length is clearly defined, a short form will be used in which all 1's are omitted (used in the tables below).

The object of this article is to compare the letter-distributions of 4-letter, 6-letter, 8-letter, 10-letter and 12-letter words under various circumstances, to find out whether the differences in the statistics are indicative of different underlying populations of words, or simply due to random fluctuations (when a coin is tossed 100 times, anywhere from 45 to 55 heads will typically occur).

What comparisons are worth making? First of all, Levine lists a host of exceedingly unusual words; is it possible that the letter-distribution of common words is different from that of rare ones? To examine this question, the letter-distributions of 4-letter through 12-letter words were tabulated with the aid of H. Kucera and W. N. Francis's Computational Analysis of Present-Day American English (Brown University Press, 1967), a tabulation of a million words published in the United States in 1961 arranged by frequency of occurrence. As shown in the table below, this represents only a small fraction of the words available in Levine.

NUMBER OF WORDS TABULATED

	4-letter	6-letter	8-letter	10-letter	12-letter
Levine	5755	28912	59540	75692	48168
Kucera-Francis	600	1000	750	600	500

The Kucera-Francis sample size was limited (for short words) by the size of the Levine corpus, and (for long words) by the necessity to sample increasingly-rare words (the rarest 12-letter words occurred only three times in a sample of a million).

A second comparison of interest is the following: is there any difference in the letter-distribution of words in dictionaries (when each different word is counted exactly once, as in Levine), and the letter-

distribution of words in text (when each word is weighted proportionally to its occurrence in text)? This question was easily answered by taking the Kucera-Francis sample and counting each word the number of times it occurred in the million words of text; thus, there were 6361 occurrences of 500 different 12-letter words, 18900 occurrences of 600 different 10-letter words, 37880 occurrences of 750 different 8-letter words, 67167 occurrences of 1000 different 6-letter words, and 112638 occurrences of 600 different 4-letter words.

A third comparison invokes the spirit of information theory, attempting to compare the actual behavior of English-language text with a theoretical model of text generation. Suppose that one places in an urn the letters of English with frequencies proportional to their occurrence in text; if "words" are formed by drawing groups of 4, 6, 8, 10 and 12 letters at a time out of this urn, will these "words" have the same letter-distribution statistics as real-life text from Kucera and Francis? To test the idea, the following distribution of 200 letters was sampled with replacement a total of 6000 times:

E 25	N 14	H 11	C 5	Y 4	V 2
T 20	I 14	D 8	M 5	P 4	K 1
A 16	S 13	L 8	U 5	W 4	JQ 1
O 16	R 12	F 5	G 4	B 3	XZ

To avoid an inordinately large sample, JQXZ was represented by a single letter and subsampled as needed.

The results of these three comparisons are presented in a series of tables on the next two pages. To facilitate comparisons, all distributions are given in percentages; 0 denotes a percentage between 0 and 0.5, whereas - denotes no occurrence of that letter-distribution in the sample. The three comparisons described above are made in columns 1-2 (Levine vs. Common List), 2-3 (Common List vs. Common Text), and 3-4 (Common Text vs. Random). The final column lists the commonest word having that letter-distribution, including the number of times it appeared in the Kucera-Francis list.

How does one interpret this mind-numbing set of tables? Which percentage differences reflect real differences between the models being compared, and which are simply statistical noise? Let us take up in turn the three comparisons introduced earlier.

A quick look at Columns 1-2 reveals some substantial differences in letter-distribution between Levine's full word list and the common word list. Nearly all of the disagreement, however, is associated with nonpattern words -- those that have all letters different. (The large disagreement in Columns 1-2 for the 11112 and 112 letter-distributions is explained by the fact that all percentages must add to 100; these distributions are the major counterweights to the nonpattern distributions.) Why is this so? The most reasonable explanation is the one hinted at earlier -- that Levine used different dictionaries to compile the nonpattern word list and the three-volume pattern word list.

PER CENT OCCURRENCE OF DIFFERENT LETTER-DISTRIBUTIONS IN WORDS OF A COMMON LENGTH

4-Letter Words

Distribution	Levine	Common List	Text	Random	
-	64	81	76	69	with-7289
2	34	18	24	29	that-10595
22	2	1	0	0	mama-44
3	0	-	-	2	lull-2

6-Letter Words

Distribution	Levine	Common List	Text	Random	
2	52	38	35	43	before-1016
-	33	51	52	39	should-888
22	10	8	11	11	people-847
3	3	3	2	5	seemed-332
23	1	0	0	1	needed-187
222	1	-	-	0	murmur-3
4	0	-	-	1	assess-6
1 other	0	-	-	0	

8-Letter Words

Distribution	Levine	Common List	Text	Random	
2	46	44	40	40	American-569
22	25	24	28	22	national-375
-	14	19	19	16	children-355
3	6	6	7	9	business-392
23	4	4	2	5	tomorrow-63
222	4	3	3	5	pressure-185
4	0	0	0	2	sessions-26
223	1	0	1	1	remember-138
33	0	0	0	0	referred-45
24	0	0	0	0	stresses-19
2222	0	-	-	0	teammate-2
6 others	0	-	-	0	

10-Letter Words

Distribution	Levine	Common List	Text	Random	
22	30	32	32	26	government-417
2	31	23	23	17	university-214
222	12	13	13	16	conditions-180
23	9	9	12	15	themselves-239
223	3	6	6	7	conference-96
3	5	6	5	6	facilities-99
-	7	5	3	3	importance-108
2222	1	2	2	2	throughout-141
24	1	1	1	2	everywhere-47
4	0	1	2	2	experience-276
224	0	1	1	1	remembered-83
33	1	0	0	1	settlement-26
233	0	1	0	1	nineteenth-42
2223	0	-	-	-	intonation-8
22222	0	-	-	-	intestines-1
12 others	0	-	-	1	

12-Letter Words

Distribution	Levine	Common List	Text	Random	
222	20	24	21	17	particularly-146
22	21	19	18	19	professional-105
223	11	13	9	12	significance-66
23	11	10	10	12	distribution-85
2222	7	10	13	6	organization-127
2	13	7	8	8	considerable-96
2223	3	5	6	5	constitution-49
3	3	4	4	4	developments-44
233	2	3	2	3	availability-21
24	2	1	2	3	nevertheless-130
224	1	1	1	2	interference-45
-	2	0	1	1	considerably-44
4	0	0	1	3	civilization-42
22222	1	1	1	-	Philadelphia-50
33	1	1	0	1	missionaries-10
234	0	1	1	1	independence-70
2233	0	0	2	1	efficiencies-98
34	1	-	-	0	hopelessness-3
2224	0	-	-	1	intermittent-3
22223	0	-	-	0	ingratiating-4
333	0	-	-	0	highlighting-2
225	0	-	-	0	dispossessed-2
235	0	-	-	-	selflessness-1
19 others	1	-	-	1	

In fact, the following table of correction factors has been derived to suggest what changes would have taken place in Levine's nonpattern word list had he used the same dictionaries that he later used for the three-volume corpus.

INFLATION FACTOR REQUIRED
TO MAKE THE NONPATTERN WORD LIST
AGREE WITH THE 3-VOLUME PATTERN WORD LIST

4 letters	6 letters	8 letters	10 letters	12 letters
2.3	2.1	1.5	0.7	0.3

These factors are entirely consistent with Levine's comment (in the preface to the nonpattern word list) that "for words of 10 or more letters an attempt has been made to give as complete a listing as practicable".

A casual comparison of Columns 2-3 reveals a high degree of agreement between them; nearly all of the observed differences can readily be explained by normal statistical variation. In other words, it is safe to conclude that there is no change in letter-distribution, whether one counts each word only once or as often as it appears in text. There are, nevertheless, a few scattered inconsistencies:

1) The 11123 distribution occurs only 17 times out of 750 in text, but in as many as 30 out of every 750 different common words -- apparently there is no outstandingly common word having this pattern, the most well-known ones being TOMORROW, EXTENDED, EXPENSES, DEMANDED and ENGINEER (all about equally likely to occur in text).
2) The 1111114 distribution occurs 12.5 times out of 600 in text, but in only 3 out of every 600 different common words -- the reason apparently being that the exceptionally common word EXPERIENCE accounts for about 90 per cent of all the text occurrences (the next commonest example is ATHABASCAN).
3) The 112233 distribution occurs 7.9 times out of 500 in text, but in only 2 out of every 500 different common words -- the reason apparently being that the very common word EFFICIENCIES accounts for about 95 per cent of all the text occurrences (the next commonest example is COMMENCEMENT).

Finally, how well does the random text model fare? There is a great deal of similarity between Columns 3-4, suggesting that it is not too unreasonable to model English text by a random mechanism as far as letter-distributions are concerned. (Parenthetically, the reader should be warned that a similar random mechanism cannot imitate the observed variation in letter-patterns for different words all having the same letter-distribution, but the demonstration of this fact is outside the scope of this article.)

Nevertheless, the samples drawn at random and from Kucera-

Francis are large enough to detect small, but apparently real, differences between the letter-distributions of randomly-generated "words" and text-generated words. Put another way, a Martian (i.e., someone from a technologically advanced civilization without the slightest knowledge of the English language) would have no difficulty in perceiving differences between "words" and words. What are they?

In general, there are too many randomly-generated "words" with letter-distributions of the form 11...1X, where X is equal to 2 or more for 4-letter words, 3 or more for 6-letter words, and 4 or more for 8-letter, 10-letter and 12-letter words. In addition, there is some evidence that there are too many randomly-generated "words" with letter-distributions of the form 11...133 and 11...124 for 8-letter, 10-letter and 12-letter words. Contrariwise, there appear to be too few randomly-generated 12-letter "words" with letter-distributions of the form 11112222 or 1122222. A genetic analogy may be helpful here. Just as a family with quadruplets or quintuplets is more likely than normal to have twins or triplets among its remaining children, a word with four or five occurrences of a single letter is likely to have more double occurrences among its remaining letters than a random model would predict.

To sum up this article: the most reliable guide to the relative frequencies of different letter-distributions is contained in Columns 2-3 based on common English words. Although the random model predicts the gross behavior correctly, it can be misleading if applied to letter-distributions which occur very rarely.